图书情报工作
第 66 卷 第 12 期 2022 年 6 月

基于共词和 Word2Vec 加权向量的文献 - 主题语义匹配分析方法*

■ 丁敬达 陈一帆 刘超 蔡微
上海大学文化遗产与信息管理学院 上海 200444

摘 要：[目的/意义]共词分析作为主题识别的重要方法，存在一定的局限和不足，将 Word2Vec 加权向量与共词分析相结合，有利于明确具体文献的主题归属，更好地对主题的发展演化进行分析。[方法/过程]在运用共词分析进行主题聚类的基础上，通过 Word2Vec 加权向量分别计算文献向量与聚类主题向量，并基于余弦相似度进行文献与主题的语义匹配。[结果/结论]国内外知识共享领域的实证分析表明，该方法能较好地将相关文献匹配至对应主题，并能从文献层面对主题特征及发展演化进行动态分析。

关键词：Word2Vec 共词分析 语义匹配 知识共享 主题演化
分类号：G203
DOI：10.13266/j.issn.0252-3116.2022.12.010

1 引言

科技文献作为一种重要的知识载体，蕴含着丰富的语义内容和主题信息，通过对海量科技文献的内容主题进行挖掘和分析，不仅有利于图书情报工作实现由文献服务、信息服务向知识服务转变，而且有助于政府、科研机构和相关人员等了解领域的主题内容、追溯主题的发展演化、把握领域的发展趋势、发现潜在的研究主题等。然而，常被用于文献主题识别研究的共词分析法存在平等对待词对共现强度[1]、无法探知具体文献所包含的主题分布[2]等不足。因此，基于 Word2Vec 能表征文本语义的特点，本研究将共词分析和 Word2Vec 结合，构建一种基于共词和 Word2Vec 加权词向量的文献 - 主题语义匹配分析方法，以便对基于共词网络的聚类主题进行发文趋势、发文时间、主题内容演变等文献层面特征的测度与分析。

2 相关研究综述

主题识别研究总体上可以分为基于引文网络的间接方法和基于文本内容挖掘的直接方法，其中，后者尤以共词分析法、LDA 主题模型等较为常见，近年来也出现了结合 Word2Vec 词向量的主题探测方法。

2.1 共词分析

共词分析是由 M. Callon 等在 20 世纪 80 年代提出[3]，利用文献集中专业词汇或者名词短语共同出现这一情况，通过大规模提取这种词语共现关系，利用聚类方法把词语和词语间复杂的共词网状关系简化为数目相对较少的类团之间的关系[4]，从而把关系不明晰的文献集中的主题直观地表达出来。从该方法被提出开始，国内外学者就围绕共词聚类分析方法的原理、应用过程等讨论了存在的问题，如忽略词间关系、忽略词在文献中的重要程度、结果独立于具体文献等[1-2,5]，并针对该方法存在的问题提出相应的改进建议或措施，如基于文献多属性加权的共词分析方法[6]、连边社团检测算法对共词分析聚类结果的改进[7]等。

2.2 结合 Word2Vec 的主题识别方法

Word2Vec 是 Google 在 2013 年开发的开源词向量训练工具，能够把文本信息从非结构化形式转化为向量化形式[8]，生成的词向量和语义相关，且更加关注上下文逻辑[9]，使得相关或者相似的词语在距离上更加接近。把 Word2Vec 等语义模型引入主题识别主要分为两种结合方式：①将主题识别模型与 Word2Vec 词向量进行模型层面的融合来提升主题识别效果，如颜端武等发现将 Word2Vec 词向量与 LDA 文档 - 主题分布

* 本文系国家社会科学基金项目“基于多元数据融合的社科领域新兴主题探测方法及实证研究”（项目编号：21BTQ010）研究成果之一。
作者简介：丁敬达，教授，博士，博士生导师，E-mail：djdhyn@126.com；陈一帆，硕士研究生；刘超，博士研究生；蔡微，硕士研究生。
收稿日期：2021-11-10 修回日期：2022-03-26 本文起止页码：108-116 本文责任编辑：易飞

相结合，能够更加全面、准确地描述微博文本的语义信息[10]；王英泽等利用Word2Vec模型将文本集转化为词汇关系矩阵，将其作为LDA模型的输入数据进行主题识别，通过对主题建模结果的解读，分析了欧盟、英国、美国颠覆性技术相关政策文本的主题特征[11]；C. E. Moody将LDA嵌入Word2Vec的学习过程中，不仅能学习单词的词嵌入，还同时学习主题表征和文档表征，提高了LDA生成主题的语义凝练度[12]；王卫军等利用Word2Vec把关键词共现关系映射到低维向量空间中，发现这种方法不仅可以完成关键词在共现网络中的重要性评价，还可以对学科关键词之间的共现关系大小进行量化[13]。②利用词向量进行文本间、词汇间相关性的匹配来实现更细化的主题分析，如闫盛枫引入Word2Vec来识别DTM主题模型下的相近主题词，实现了主题词中同义词的归并[14]；周云泽等选取LDA所识别主题中隶属概率最高的10个主题词与Word2Vec词向量相结合的方法来表征主题向量，以实现相似主题的匹配[15]；C. Li等也证明了Word2Vec与LDA模型结合加权向量能够有效将技术主题特征表示为低维稠密的向量形式，并利用余弦相似度实现了文献和主题在语义上的匹配[16]，从而实现更为精细化的语义建模。

综上所述，Word2Vec与主题词、文本词汇的有机结合，可以有效地表征主题或者文本的语义特征，实现更细粒度的语义关联与分析，目前研究主要聚焦于将LDA与Word2Vec相结合，但LDA主题模型更适用于长文本，对一些短文本的主题识别效果不佳，此外模型的主题数目也需要根据困惑度曲线人为确定，针对这些问题虽有一些改进[17-18]，但总体还没有形成相对成熟的措施。此外，将共词分析与Word2Vec相结合的研究相对较少，主要是利用Word2Vec学习或者替代共现关系[13,19]，较少利用到Word2Vec在文本匹配上的优势。考虑到共词分析所得主题仅表现为不同关键词的聚类，其结果独立于文献，为克服对于任意一篇文献无法探知其中所包含主题分布的不足[2]，本研究尝试将共词分析和加权的Word2Vec结合，利用Word2Vec能表征文本语义的特点，构建一种基于共词分析和Word2Vec加权词向量的主题－文献语义匹配分析方法，并用于对主题进行文献层面的特征测度及其发展脉络的演化分析。

3 方法构建和主题测度

3.1 方法构建

为解决共词分析无法进行文献层面计量的问题，本研究将Word2Vec模型应用于共词分析，实现共词网络下的主题－文献相似度匹配，从而将不同文献划分给对应的主题。首先，利用题录数据提供的关键词信息构建分词所需的领域词典，在数据清洗后选取高频词构建关键词共现网络并进行主题聚类；其次，利用题目、摘要和关键词作为文本数据训练Word2vec词向量，基于词向量构建主题向量和文献向量；最后，根据设定的规则实现文献与主题的匹配，并选取主题测度指标对结果进行测度与分析。具体流程如下：

3.1.1 构建共词矩阵

利用Python的jieba分词包，提取题录数据中的关键字建立分词所需的领域词典，进行分词和词性筛选，在此基础上构建共词矩阵，分为三个步骤：①同义词归并；②高频关键词选取；③矩阵构建。其中，对于同义词的归并，采用$(word_i \cap word_j)/(word_i \cup word_j)$计算关键词之间的相同字符重叠度，$word_i$表示关键字$i$中的字符集合，如关键字“结构方程模型”的字符集合为{结，构，方，程，模，型}，针对相同字符重叠度较高的词汇再辅以人工筛选得到同义词；对于高频关键词的筛选，本研究采用普赖斯公式$M = 0.749\sqrt{N_{max}}$[20]，其中N_{max}为词频最高的关键词出现次数。对共词矩阵进行聚类，从中提取出不同研究主题，每个主题下的关键词就是该主题的主题词。

3.1.2 进行文献－主题匹配

Word2Vec词向量模型本质上是具有“输入层－隐藏层－输出层”的三层神经网络模型[8]，如图1所示，w(t)为目标词，其上下文词汇为w(t－r)、…、w(t－1)、w(t+1)、…、w(t+r)。该模型有CBOW(Continuous Bag of Words)和Skip-gram两种学习方式，其中，Skip-gram模型是根据目标词预测目标词的上下文。本研究采用Skip-gram学习方式，利用题目、摘要和关键词的组合来代替单篇文献，将样本数据作为训练Word2Vec模型的数据集。然后对共词矩阵聚类得到的主题$topic_t$，利用训练好的Word2Vec模型生成$topic_t$中每个主题词的词向量，基于词频关系将词向量进行加权求和从而得到$topic_t$的向量化表示：

$$W_{topic_t} = w_1 \times w2v_{k_t^1} + w_2 \times w2v_{k_t^2} + \cdots + w_{k_t} + w2v_{k_t^{k_t}}$$

$t = 1,2,\cdots,T$ 公式(1)

$$w_i = \frac{Freq(k_t^i)}{\sum_{j=1}^{k} Freq(k_t^j)} \quad i = 1,2,\cdots,k_t$$ 公式(2)

其中，$w2v_{k_t^i}$指代主题词k_t^i的Word2Vec词向量，T代表主题数，k_t表示$topic_t$的主题词数量，w_i是k_t^i词向

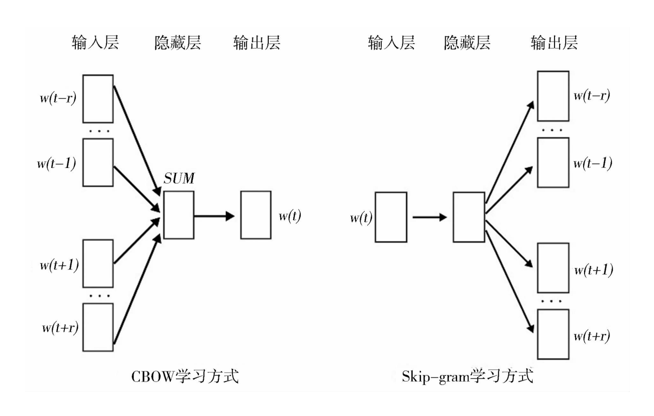


图 1 CBOW 和 Skip-gram 模型网络结构[8]

量的权重，即主题词 k_t^i 的频次与该主题下所有主题词的总频次之比。

在文献向量化表示的处理上，依然采用题目、摘要和关键词的组合（代替单篇文献）作为数据源，但摘要中可能会存在一些高频的无关词语，为了更好地衡量词汇的重要程度，我们采用 TF-IDF 对数据源中每个词语的词向量进行加权进而得到文献向量 $w2v_tfidf_{d_i}$，以降低区分度低的高频词的影响力[21]和提高 Word2Vec 的特征表示效果[22]。最后，通过余弦相似度计算文献向量 $w2v_tfidf_{d_i}$ 与主题向量 W_{topic_t} 之间的相关性，得到每篇文献和各个主题之间的相似程度：

$$Similar_{topic_t d_i} = cosine(W_{topic_t},\ w2v_tfidf_{d_i}) \quad \text{公式(3)}$$

对于 T 个主题、D 篇文献来说，总共需要计算 $T \times D$ 次，文献隶属主题匹配规则为：①如果某文献 d_i 对于任意一个主题 $topic_t$ 的 $Similar_{topic_t d_i}$ 大于等于 β，则该文献隶属于 $topic_t$；②如果某文献 d_i 对于任意一个主题 $topic_t$ 的 $Similar_{topic_t d_i}$ 都小于 β，则认为该文献仅隶属于最大 $Similar_{topic_t d_i}$ 的 $topic_t$。通过这种方法，可以把一篇文献分配到不同的主题，且每篇文献可能不止对应一个主题，这也和实际情况相符合，因为很多文献的内容可能会和多个主题相关。

传统上共词分析可以结合社会网络分析、多维尺度分析、战略坐标图等方法来对主题进行识别，但都依赖关键词进行分析，相应的主题演化分析也只是对不同时间段聚类主题进行分析，究其原因在于关键词和文献之间并无连接，本方法实现了共词网络下的主题－文献之间的匹配，使得共词分析能从文献层面量化研究主题的演化脉络。此外，现有基于 Word2vec 构建主题向量的方法往往固定选择每个主题前 h 个主题词，没有考虑不同主题规模不同所带来的影响，本方法基于共词网络的高频关键词构建主题向量则更全面包含了主题的语义特征。

3.2 主题测度和演化分析

3.2.1 主题特征测度

对于主题特征的测度指标，常见的有主题强度[23]、新颖度[24-25]、影响力[26]、交叉性[27]、关注度[28]等，总的来看，相关的测度指标通常会引入主题对应的文献数量以及发表时间，因此本研究采用主题强度、关注度、新颖度三项指标研究各个主题的特点。

（1）主题强度（*Strength Index*，*SI*）。主题强度是一个主题热门与否的最直观的表现。从数量上看，一个主题累积的文献越多，说明科研人员对其投入的精力越大，在学术领域中的影响力越深远，该主题的强度也越强：

$$SI = \sum_{i=1}^{D} d_{topic_t}^i \quad \text{公式(4)}$$

其中，$d_{topic_t}^i = \begin{cases} 1 & \text{如果文献 } i \text{ 隶属于主题 } t \\ 0 & \text{如果文献 } i \text{ 不属于主题 } t \end{cases}$，其中主题－文献匹配方法同上。

（2）关注度（*Attention Index*，*AI*）。关注度是一个动态变化的过程，需要从时间和数量两方面进行描述。从时间维度上看，由于科研人员的注意力有限，再加上主题自身发展状况与社会发展变化等因素，科研人员对于某个主题的关注程度随时间会产生波动；从数量上看，关注度相当于每年的主题强度高低，即每年该主题下的文献数量，利用主题－文献匹配方法获取各主题下每年产出的文献数量，可以量化科研人员对于主题的关注程度：

$$AI = SI_{topic_t}^{year} \quad \text{公式(5)}$$

其中，$SI_{topic_t}^{year}$ 代表每年隶属于 $topic_t$ 的文献数量，其中主题－文献匹配方法同上。

（3）新颖度。由于文献随着“年龄”的增长，其内容会日益变得陈旧过时，作为情报源的价值不断降低，而新文献的涌入，伴随着可能带来的新的理论、方法和观点等，也会加快原有文献价值的衰减，因此，通过主题－文献匹配方法获取主题下对应文献后，可以进一步测度这些文献的新颖度并将其作为判断该主题发展潜力的一项重要指标。文献首次公开发表年份是揭示文献新旧的常见指标，一个主题的新颖程度可以用隶属该主题文献发表年份的中位数表示，中位数越大代表了该主题内的大部分文献出版年份越靠前，出现新成果的可能性越高。

3.2.2 主题演化分析

相比于利用传统共词分析分阶段研究主题演化的方式，本研究可以从时间维度对于主题发展脉络进行

直接分析，具体方法为：在利用上述方法获取每个研究主题对应文献的前提下，把主题下的文献按年划分，将每年的文献关键词作为语料库来计算不同年份间关键词的 TF-IDF 值。按照 TF-IDF 值降序排列，可以获得每个主题每年的核心关键词，通过关键词突现分析可以对于主题的研究脉络有一个动态的宏观认识，辅以该主题下具有对应关键词的文献的内容分析能够较为细致的了解该主题的发展与演变。

4 实证分析

随着知识经济全球化的到来，知识的生产、加工、创新和应用日益成为推动经济增长和社会发展的主导力量，无论是对企业组织或是个人，知识都被视为关键的战略资源[29]，而知识共享作为分享、利用和创造知识的关键过程更是受到了企业、学者等多方关注，但与日俱增的文献使人们难以把握其核心知识，为较为全面地把握国内外知识共享研究领域的知识体系和发展前沿、找准研究的切入点、提升企业竞争力，本研究以“知识共享”领域为例对文献 – 主题匹配方法进行实证分析，在此基础上测度知识共享主题特征并进行演化分析。

4.1 数据来源及处理

中文数据来源于 CNKI，以“知识共享”为主题进行检索，来源数据限定为核心期刊、CSSCI、CSCD，检索时间为 2021 年 4 月 20 日，共得到期刊论文 5 481 篇，去除公告、报道等无关数据后，得到期刊论文 5 132 篇；英文数据来源于 Web of Science 核心合集，以“knowledge sharing”为主题进行检索，限定语种为 English，时间跨度为 1996 年 – 2020 年(受限于数据库的使用权限，本单位 IP 仅能检索 1996 年及之后的数据。知识共享的研究起源于 1990 年，但 1990 – 1995 年的发文量很少[30])，检索时间为 2021 年 4 月 20 日，共得到期刊论文 5 813 篇，在去除无关文献后，得到期刊论文 5 625篇。每年文献分布见图 2，国内论文数量在 2010 年后逐年下降，但国外论文数量在 2015 年突然激增，并在此后每年都持续上升。

4.2 共词聚类

在构建关键词共现网络过程中，由于“知识共享”是本文研究的主题词，“知识管理”的含义较为宽泛，且出现频率过高不利于聚类，故在后续研究中去除这两个词语，选取剩余关键词进行清洗，经过多次试验，选取重叠度在 0.6 以上的关键词进行人工筛选后实现同义词归并，后根据普赖斯公式计算高频关键词阈值，

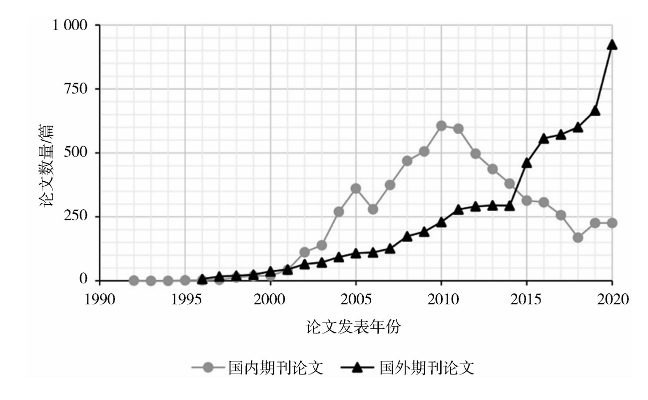


图 2 知识共享领域国内外文献年度分布

选定出现次数在 14(国内)、11(国外)以上的词作为高频关键词，其中国内论文包含 125 个、国外论文包含 277 个。将关键词共现矩阵进行主题聚类，其聚类结果强度分布如图 3 所示：

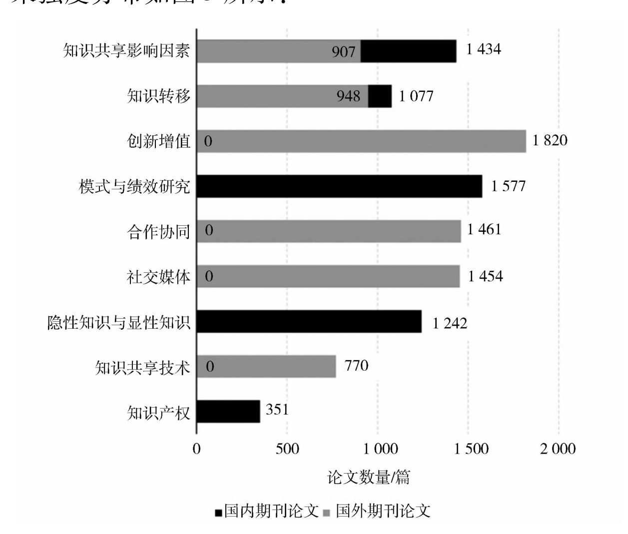


图 3 国内外知识共享研究主题强度分布

4.3 文献 – 主题语义匹配

通过 Python 的 jieba、nltk 库进行数据处理，“jieba.load_userdict()”加载自定义领域词典，使用 pytorch 实现 Word2Vec 词向量的训练模型搭建。然后，求出每个关键字的词频以及 TF-IDF 值，结合训练好的词向量加权求和分别得到主题向量和文献向量，利用余弦相似度进行主题向量与文献向量的匹配。经过多次实验后，发现将国内文献匹配阈值取值为 0.62、国外文献匹配阈值取值为 0.24 时文献 – 主题划分效果较好。表 1 – 表 4 列举了国内“隐性知识与显性知识”主题和国外“社交媒体”主题下 10 篇最高相似度与 10 篇最低相似度的文献，发现和主题向量具有较高相似度的文献的标题往往包含一些核心关键词，而与主题相似度较低的文献则大多可以作为该主题的延伸性研究。

表 1 "隐性知识与显性知识"主题下相似度最高的 10 篇文献

文献标题	相似度
基于知识管理的隐性知识显性化研究	0.796
高校科研团队个体成员隐性知识隐藏意愿分析	0.796
高校科研团队隐性知识共享意愿影响因素研究——中国文化情境下基于使能和抑制的双重视角	0.796
科研团队显性知识和隐性知识共享意愿影响因素的对比分析	0.795
基于 SD 模型的虚拟社区中个体隐性知识共享行为研究	0.795
企业隐性知识沟通的动力机制研究	0.794
组织文化变革中的隐性知识转移研究	0.794
隐性知识创新影响因素的实证研究	0.794
知识管理在图书馆中的实施	0.792
知识型企业人际信任倾向对员工隐性知识共享影响的实证研究	0.792

表 2 "隐性知识与显性知识"主题下相似度最低的 10 篇文献

文献标题	相似度
知识管理在高职院校教学基本建设中的应用	0.612
浅析知识管理背景下的高校档案馆管理	0.612
知识管理与图书馆可持续发展	0.612
企业隐性知识显性化过程与机制研究	0.611
企业内部隐性知识如何转化为显性知识?——基于国企的案例研究	0.611
聆听与分享:真人图书馆在中国的实践及思考	0.611
知识科学视角下我国知识融合研究现状解析	0.341
利用信息技术整合企业培训资源的路径选择	0.340
高校智力资本模型及实证研究	0.337
基于知识地图的 MOOC 学习共同体的学习研究	0.332

表 3 "社交媒体"主题下相似度最高的 10 篇文献

文献标题	相似度
What factors influence knowledge sharing in organizations? a social dilemma perspective of social media communication	0.556
Social-media-based knowledge sharing: a qualitative analysis of multiple cases	0.556
Can lean media support knowledge sharing? investigating a hidden advantage of process improvement	0.553
Study of social media impacts on social capital and employee performance - evidence from tunisia telecom	0.548
How do features of social media influence knowledge sharing? an ambient awareness perspective	0.547
The use of social media in knowledge sharing case study undergraduate students in major british universities	0.547
The role of social identity and communities of practice in mergers and acquisitions	0.544
Dynamic competition strategy for online knowledge-sharing platforms	0.535
To share or hide? a social network approach to understanding knowledge sharing and hiding in organizational work teams	0.533
Users' knowledge sharing on social networking sites	0.533

表 4 "社交媒体"主题下相似度最低的 10 篇文献

文献标题	相似度
Social networks under stress: specialized team roles and their communication structure	0.231
Intentionally creating a community of practice to connect dispersed technical professionals	0.231
Knowledge-based network participation in destination and event marketing: a hospitality scenario analysis perspective	0.230
Enacting knowledge strategy through social media: passable trust and the paradox of nonwork interactions	0.230
Negotiating the expertise paradox in new mothers' whatsapp group interactions	0.226
Knowledge-sharing networks in hunter-gatherers and the evolution of cumulative culture	0.225
Gamifying knowledge sharing in humanitarian: a design science journey	0.218
Online formative assessments with social network awareness	0.214
Virtual knowledge brokering: describing the roles and strategies used by knowledge brokers in a pediatric physiotherapy virtual community of practice	0.212
Library and information science's ontological position in the networked society: using new technology to get back to an old practice	0.180

4.4 主题特征分析

4.4.1 主题强度与关注度

国内外研究主题强度分布如图 3 所示,国内研究主题聚焦在知识共享影响因素、知识共享模式与绩效研究方面;国外把研究的重点放在了知识共享带来的创新增值以及组织间、个人间的合作协同研究。相比于国内外知识共享主题关注度变化情况(见图 4),对于国内研究来说,除影响因素主题外,其他主题的研究都在 2010 年前后开始下降;如 2009 年之前知识共享模式与绩效研究的关注度高于影响因素,但其相关研究在 2009 年后关注度逐年下降,而知识共享影响因素的相关研究受到了更多的重视,每年都保持着较高且稳定的成果数量,成为了当下研究的主流;对于国外研究来说,由于国外知识共享还处在成果涌现的成长期,导致了强度大小与关注度程度呈现了相似的变化,如 2007 年以来,基于知识共享带来创新增值的研究受到

的关注逐渐增强,发展为当前研究主流,而合作协同的研究也在 2017 年一举超越社交媒体,成为了国外科研人员第二大关注的主题。

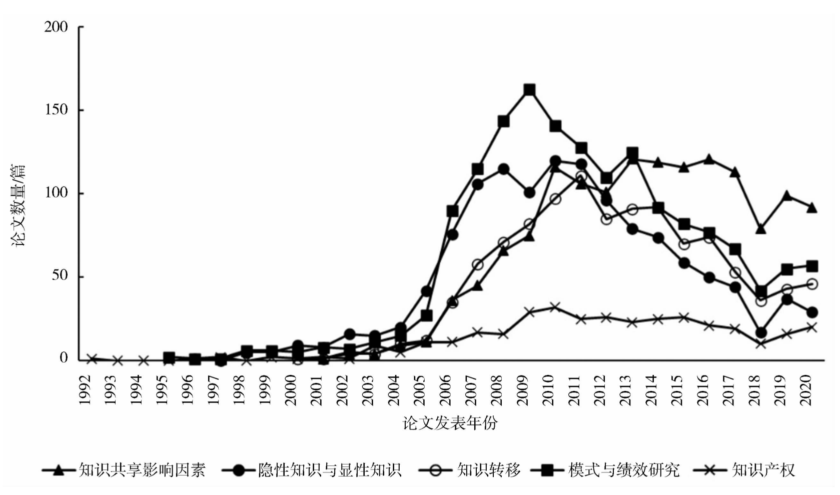


图 4　国内研究热点主题关注度变化

从国内外知识共享研究的主题重叠与独立程度来看,知识共享影响因素、知识转移这两个主题在国内外研究中都有凸显,但二者的关注度又有所不同。国内外对于影响因素的关注度比较高,而相比于国外,国内科研人员对于知识转移的关注自 2011 年起逐年下降(见图 4 和图 5)。国内隐性知识与显性知识、知识共享模式与绩效研究、知识产权研究以及国外基于社交媒体、创新增值、知识共享技术以及合作协同的研究展现了国内外知识共享研究的不同发展方向。对于国内研究来讲,科研人员对于隐性知识与显性知识、知识共享模式与绩效研究的关注度在 2008－2010 年达到峰值后开始逐年下降,而对于知识产权的关注总体上呈现比较平稳的态势;对于国外研究来讲,基于社交媒体、创新增值以及合作协同关注度总体呈上升趋势,而知识共享技术关注程度每年都比较平稳。

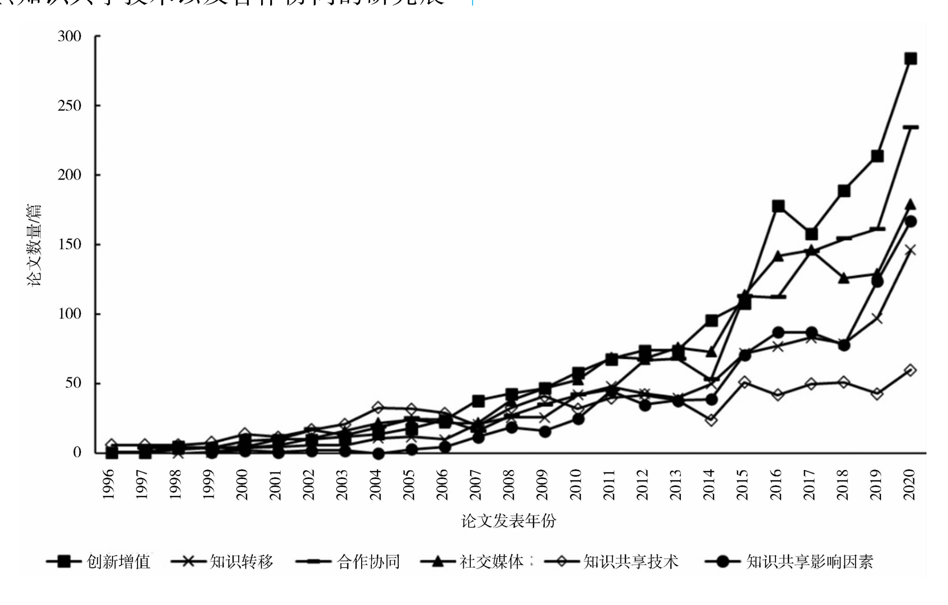


图 5　国外研究热点主题关注度变化

总体上,相对于国内知识共享研究,国外更加注重技术迭代、组织间合作、知识共享带来的知识、产品、服务的创新增值以及基于社交媒体的知识共享,且研究的递进性较强,表现为关注度的上升有一个缓慢积累到快速攀升的过程。而国内知识共享研究的内容更注重影响因素及各类应用,表现为关注度的起伏变化更大,对知识共享技术和社交媒体下知识共享的研究重视不足。

4.4.2　主题新颖度

国内外每个研究主题对应文献的出版年份分布如

图 6 所示。国外箱型图的中位线整体高于国内箱型图的中位线，由此可知，国外对于知识共享研究的新颖度整体高于国内。具体来看，国内知识共享主题新颖度主要分布在 2010 – 2013 年，知识共享的影响因素以及跨组织、团队之间的知识转移这两个主题内文献的发表年份比较新，知识共享模式与绩效研究、知识产权的研究虽然起步早，但伴随着新技术的引入、相关法律法规的出台，新研究主题不断涌现；国外知识共享主题新颖度主要分布在 2011 – 2014 年间，知识共享影响因素、合作协同近年来受到了科研人员的持续关注，知识共享技术的新颖度较低可能是因为相关技术已经比较成熟，研究成果正逐步运用到其他主题之中。

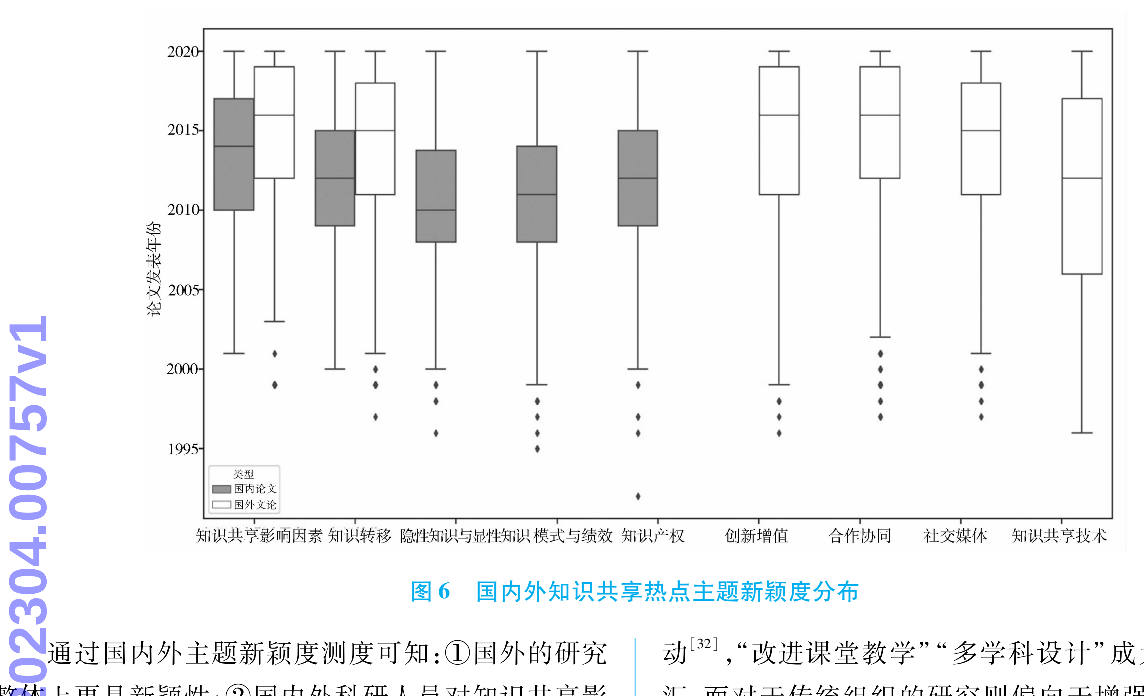


图 6　国内外知识共享热点主题新颖度分布

通过国内外主题新颖度测度可知：①国外的研究整体上更具新颖性；②国内外科研人员对知识共享影响因素的研究与运用在不断地更新、进步，且个人、组织间的知识转移以及合作协调也依然保持着较高的被关注度；③主题研究起点的早晚和主题新颖度之间没有直接联系；④国内对于影响因素的研究与国外对于创新增值、合作协同的研究的强度、新颖度都很高，较可能产生比较前沿的研究方向或发展趋势。

4.5　主题演化分析

图 7 绘制了“合作协同”主题下国外科研人员关注情况，并列出部分年份 TOP5 关键词。国外对于合作协同的研究，主要聚焦在个人、组织、群体间如何实现协同来达到高效的知识共享，随时间推进，研究对象从企业逐渐向城市、国家、虚拟社区扩散，研究的问题也更加多元化。主题发展初期主要以系统平台为媒介研究企业间的战略合作以及知识共享实践面临的问题，基于本体方法的“多智体系统”成为了突现的主题词，通过建立完善的智体系统可以实现组织内部各个环节之间的协同运作，过程的优化改进也有助于知识共享的实现[31]；2007 年，知识共享理论与技术被引入高等教育领域并受到了科研人员的关注，在传统的课堂教学中利用信息和通信技术来提升课堂协作与群体互动[32]，“改进课堂教学”“多学科设计”成为了核心词汇，而对于传统组织的研究则偏向于增强“利益相关者”间的知识共享来实现组织的高效运转；2011 年起，利用知识共享促进城市[33]、公共部门的可持续发展的研究开始兴起，另一方面，随着气候问题的日益严重，把知识共享带入气候变化适应研究中将促进决策的执行和对突发情况的应对，如针对不同国家治理气候成功案例的学习以及所用理论框架、数据的共享[34]；此外，通过知识共享使医生、患者多方协同参与治疗与护理[35]在 2016 年成为一个新颖的研究方向。2020 年，受到新冠疫情的影响，组织的工作模式与环境发生了很大的变化，在线方式下如何提高学生、办公人员、医疗人员知识共享需要进一步探讨与实践，社交媒体与虚拟社区中针对疫情相关消息与防疫知识的共享[36-37]受到了科研人员的广泛关注。

5　结语

本文提出了一种 Word2Vec 加权向量和共词分析相结合的文献 – 主题匹配分析方法，并以国内外知识共享领域为例进行实证分析，以弥补共词分析在文献层面测度的不足。首先，采用自然语言处理与文本挖掘技术对国内外知识共享文献的题录数据进行了清

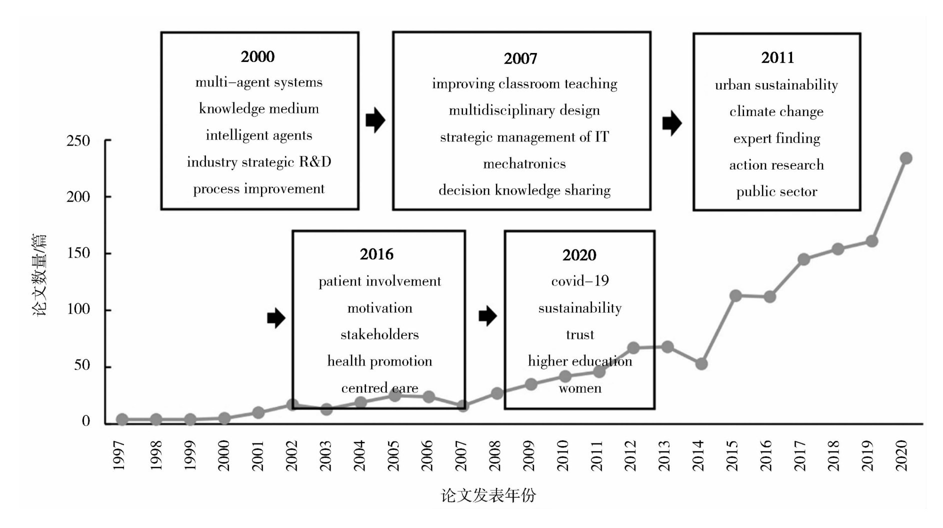


图7 "合作协同"研究主题的演化

洗,其次通过共词分析得到国内外知识共享领域的相关研究主题,然后利用加权 Word2Vec 词向量把文献和相应研究主题进行匹配。实证分析结果表明,该方法能够获取与研究主题高度相关的文献;相比于传统共词分析,该方法不仅能从宏观上探测主题的演化过程,也能利用现有的主题测度指标从文献这一角度评价主题的发展状况,结合主题词突现深入剖析主题的发展脉络与动态演变。本研究局限性在于:这是一种无监督的方法,阈值需要根据匹配结果进行主观调整,较高的阈值虽然可以提高主题对应文献的准确性,但也会导致一些文献的潜在主题被忽略,在未来可以参考监督主题模型的思想,如 Label-LDA、MedLDA 等方法,结合出版地、作者等可观察到的文献外部特征信息对文献进行标注,以实现最优阈值的自动化生成。

参考文献:

[1] 巴志超,李纲,朱世伟. 共现分析中的关键词选择与语义度量方法研究[J]. 情报学报,2016,35(2):197 - 207.

[2] 周利琴,徐健,巴志超,等. 基于 SNA 和 DMR 方法的高血压主题探测与演化趋势比较研究[J]. 图书情报工作,2018,62(13):82 - 91.

[3] CALLON M, COURTIAL J P, TURNER W A, et al. From translations to problematic networks: an introduction to co-word analysis[J]. Social science information,1983,22(2):191 - 235.

[4] 钟伟金,李佳,杨兴菊. 共词分析法研究(三)——共词聚类分析法的原理与特点[J]. 情报杂志,2008(7): 118 - 120.

[5] 李纲,巴志超. 共词分析过程中的若干问题研究[J]. 中国图书馆学报,2017,43(4):93 - 113.

[6] 李锋. 基于核心关键词的聚类分析——兼论共词聚类分析的不足[J]. 情报科学,2017,35(8): 68 - 71,78.

[7] 孙海生. 连边社团检测算法对共词分析聚类结果的改进研究[J]. 图书情报工作,2016,60(10):123 - 129.

[8] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[EB/OL]. [2022 - 03 - 22]. https://arxiv.org/abs/1310.4546v1.

[9] 裘惠麟,邵波. 多源数据环境下科研热点识别方法研究[J]. 图书情报工作,2020,64(5):78 - 88.

[10] 颜端武,梅喜瑞,杨雄飞,等. 基于主题模型和词向量融合的微博文本主题聚类研究[J]. 现代情报,2021,41(10):67 - 74.

[11] 王英泽,化柏林. 欧美国家颠覆性技术政策文本数据的主题建模分析研究[J/OL]. 情报理论与实践,2022:1 - 14 [2022 - 03 - 22]. http://kns.cnki.net/kcms/detail/11.1762.g3.20220225.1702.002.html.

[12] MOODY C E. Mixing dirichlet topic models and word embeddings to make lda2vec[J/OL]. [2022 - 03 - 23]. https://arxiv.org/abs/1605.02019.

[13] 王卫军,姚畅,乔子越,等. 基于词嵌入的国家自然科学基金学科交叉知识发现方法——以"人工智能"与"信息管理"为例[J]. 情报学报,2021,40(8):831 - 845.

[14] 闫盛枫. 融合词向量语义增强和 DTM 模型的公共政策文本时序建模与演化分析——以"大数据领域"为例[J]. 情报科学,2021,39(9):146 - 154.

[15] 周云泽,闵超. 基于 LDA 模型与共享语义空间的新兴技术识别——以自动驾驶汽车为例[J/OL]. 数据分析与知识发现,2021:1 - 16 [2022 - 03 - 25]. http://kns.cnki.net/kcms/detail/10.1478.g2.20211206.1917.007.html.

[16] LI C, GUO J, LU Y, et al. LDA meets Word2Vec: a novel model for academic abstract clustering[C]//Companion of the Web Conference 2018. Republic and Canton of Geneva: CHE, 2018:1699 - 1706.

[17] 蒋甜,刘小平,刘会洲. 基于关键词关联度指标(KRI)进行 LDA 噪声主题过滤的方法研究[J]. 图书情报工作,2020,64(3):92 - 99.

[18] 王婷婷,韩满,王宇. LDA 模型的优化及其主题数量选择研究——以科技文献为例[J]. 数据分析与知识发现,2018,2(1):29 - 40.

[19] HUANG L, CHEN X, ZHANG Y, et al. Identification of topic evolution: network analytics with piecewise linear representation and

word embedding[J]. Scientometrics, 2022, 127(2):1-31.
[20] 虞秋雨，徐跃权. 共词分析中高频词阈值确定方法的实证研究——以新冠肺炎文献高频词选取为例[J]. 情报科学, 2020, 38(9):90-95.
[21] 白如江，刘博文，冷伏海. 基于多维指标的未来新兴科学研究前沿识别研究[J]. 情报学报, 2020, 39(7):747-760.
[22] TANG X, WAN Y, LIU Y, et al. Chinese spam classification based on weighted distributed characteristic[C]//Proceedings of the 2017 Chinese Automation Congress. Jinan: 2017: 6618-6622.
[23] 白敬毅，颜端武，陈琼. 基于主题模型和曲线拟合的新兴主题趋势预测研究[J]. 情报理论与实践, 2020, 43(7):130-136, 193.
[24] 吴一平，于纯良，曲佳彬，等. 文本主题视域下的高校论文研究前沿领域及演化发展趋势研究[J]. 情报科学, 2021, 39(5):156-162,183.
[25] 黄璐，朱一鹤，张嶷. 基于加权网络链路预测的新兴技术主题识别研究[J]. 情报学报, 2019, 38(4):335-341.
[26] 范少萍，安新颖，晏归来，等. 医学领域前沿主题识别方法研究[J]. 情报学报, 2018, 37(7):686-694.
[27] 刘自强，许海云，岳丽欣，等. 面向研究前沿预测的主题扩散演化滞后效应研究[J]. 情报学报, 2018, 37(10):979-988.
[28] 熊回香，李跃艳. 基于 Word2vec 的科研人员推荐与跨语言论文推荐模型研究[J]. 情报科学, 2019, 37(12):19-26.
[29] CASTANEDA D I, CUELLAR S. Knowledge sharing and innovation: a systematic review[J]. Knowledge and process management, 2020, 27(3):159-173.
[30] 张春阳，梁启华. 基于 Web of Science 知识共享科学研究现状及发展态势分析[J]. 图书馆学研究, 2016(18):20-29.
[31] KOCK N, DAVISON R. Can lean media support knowledge sharing? investigating a hidden advantage of process improvement[J]. IEEE transactions on engineering management, 2003, 50(2):151-163.
[32] LOOI C K, CHEN W. Community-based individual knowledge construction in the classroom: a process-oriented account[J]. Journal of computer assisted learning, 2010, 26(3):202-213.
[33] SHEN L Y, OCHOA J J, SHAH M N, et al. The application of urban sustainability indicators - a comparison between various practices[J]. Habitat international, 2011, 35(1):17-29.
[34] JOHANNA M, NATASHA K, ARNOLDO M K, et al. Climate adaptation research for the next generation[EB/OL]. Climate and development, 2013:189-193 [2022-03-25]. https://www.tandfonline.com/doi/full/10.1080/17565529.2013.812953.
[35] GEORGIA T BN, TRACEY B, ANDREA M, et al. Patients' perceptions of participation in nursing care on medical wards[J]. Scandinavian journal of caring science, 2016, 30(2):260-270.
[36] EDGHIEM F, ABUALQUMBOZ M, MOUZUGHI Y. Covid-19 transition, could Twitter support UK Universities? [J/OL]. Knowledge management research & practice, 2020:1-6 [2022-03-25]. https://www.tandfonline.com/doi/full/10.1080/14778238.2020.1848364.
[37] SAKUSIC A, MARKOTIC D, DONG Y, et al. Rapid, multimodal, critical care knowledge-sharing platform for COVID-19 pandemics[J]. Bosnian journal of basic medical sciences, 2020, 21(1):93-97.

作者贡献说明：
丁敬达：论文选题、研究思路和框架制定，论文撰写；
陈一帆：数据处理和分析，研究思路和框架制定，论文撰写；
刘超：研究思路和框架制定；
蔡微：数据采集和分析。

An Article-Topic Semantic Matching Analysis Method Based on Co-Word and Weighted Word2Vec

Ding Jingda Chen Yifan Liu Chao Cai Wei
School of Cultural Heritage and Information Management, Shanghai University, Shanghai 200444

Abstract: [**Purpose/Significance**] As an important method for topic identification, co-word analysis has some limitations and deficiencies. The combination of weighted Word2Vec and co-word analysis is helpful to clarify the topic attribution of specific articles, and to better analyze the evolution of topics. [**Method/Process**] On the basis of topic clustering by co-word analysis, the article vectors and the clustering topic vectors were calculated by weighted Word2Vec, and the semantic matching between articles and topics was carried out based on cosine similarity. [**Result/Conclusion**] The empirical analysis in the field of knowledge sharing at home and abroad shows that this method can better match the relevant articles to the corresponding topics, and a dynamic analysis of the topic characteristic and evolution can be carried out from the article level.

Keywords: Word2Vec co-word analysis semantic matching knowledge sharing topic evolution